PHYSICAL-CHEMICAL PROPERTY BASED SEQUENCE MOTIFS AND METHODS REGARDING SAME

Cross-Reference to Related Applications

This application claims the benefit of U.S. Provisional Application No. 60/460,769, entitled "PHYSICAL-CHEMICAL PROPERTY BASED SEQUENCE MOTIFS AND METHODS REGARDING SAME," filed 4 April 2003, wherein such document is incorporated herein by reference.

Background of the Invention

The present invention relates generally to the analysis of sequence data. More particularly, the present invention pertains to defining physical-chemical property (PCP) based sequence motifs common to a family of related proteins and/or the use thereof in analysis of sequence data (e.g., DNA, RNA, amino acids), such as, for example, searching genomic sequence databases to identify homologues of the family of proteins.

With improvement of technology, the amount of sequence data available for analysis is accumulating very quickly. The ability to analyze such sequence data depends significantly on the development of advanced computational tools for rapid and accurate annotation of genomic sequences as to the probable structure and function of the proteins they encode.

As such, one of the most challenging goals of genome sequencing projects is to functionally annotate novel gene products (Kelley, et al., 2000 and Rison, et al., 2000). A sequence can be recognized as a homologue of a known protein if the pair-wise sequence identity/similarity exceeds a statistically derived threshold (e.g., more than 30% sequence identity or an E-value less than 0.001) (Chothia, et al., 1986). These global criteria identify only a small fraction of proteins known to be functionally related, as amino acids patterns are differently conserved.

Determining the similarity of sequences in databases to that of proteins of known function has been one of the most direct computational ways of deciphering codes that connect molecular sequences of protein structure and function. There are various algorithms and software available for sequence database searching and sequence analysis which, for

example, may provide for comparisons between query sequences and sequence data (e.g., sequence data in a molecular database). Sequence profile searches (Bowie, et al., 1991; Gribskov, et al., 1996; Mehta, et al., 1999; Rychlewski, et al., 2000; and Schaffer, et al., 1999) and Hidden Markov Models (Eddy, S.R., 1998) generate position specific fingerprints of the amino acid sequences in protein families and can identify distantly related proteins. However, the optimal choice of parameters for high sensitivity/specificity depends on the expert user. A further complication is that enzymes often combine functional elements to create a specific catalytic center. These elements, due to crossover events, may not occur in the same linear fashion in the sequence of related proteins and are not found with global profiles.

Analytical tools that use statistically derived matrices based on allowed substitution of amino acids, are not designed to detect conservation of physical-chemical properties. For example, such tools include FASTA, PSI-BLAST, or BLOCKS, such as described in Pearson W., Rapid and sensitive sequence comparison with FASTP and FASTA, *Methods in Enzymology*, 1990, 183:63-98; Schaffer et al., Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, *Nucleic Acids Res*, 2001, 29(14):2994-3005; Schaffer et al., IMPALA: matching a protein sequence against a collection of PSI-BLAST constructed position-specific score matrices, *Bioinformatics*, 1999, 15:1000-1011; Altschul et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, 1997, 25(17):3389-3402; and Henikoff et al., Increased converage of protein families with the blocks database servers, *Nucleic Acids Res*, 2000, 28:228-230.

Other database searching tools that are based on information derived from a family of related proteins include, for example, a screening for motif patterns such as described in U.S. Patent No. 5,845,049, to Wu, issued December 1, 1998, and entitled "Neural network system with n-gram term waiting method for molecular sequence classification and motif identification." Wu describes a method using a neural network that is trained for extraction of sequence motifs. Further, for example, other analysis processes may use other fold recognition tools (e.g., U.S. Patent No. 6,512,981 B1, to Eisenberg, et al., issued January 28, 2003, entitled "Protein fold recognition using sequence-derived predictions"). Eisenberg et al. describes a method that relies, to a large extent, on the knowledge of 3D protein structures

for fold assignment.

Most genome sequencing projects represent their results in databases including large collections of sequences. As such, a critical step in the selection of potential drug targets among novel gene products is functional annotation. Global sequence similarity criteria can only identify a small fraction of proteins known to be functionally related, as amino acid patterns are not uniformly conserved and it is not known what physical-chemical properties are conserved. The large number of potential physical-chemical properties makes it difficult to know 'a priori' which of these properties and at what positions in the protein sequence these properties are conserved. A process for deriving five descriptors for amino acids using 237 physical-chemical properties is described in the article by Venkatarajan, M.S. and Braun, W. (2001), entitled "New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties," *J Mol. Model.*, **7**, pp 445-453.

The available or described sequence data analysis methods range from very sensitive, but computationally intensive algorithms, to relatively rapid, but less sensitive analysis methods. As such, although various analysis tools are available, there is still a need for the development of database search methods that are relatively rapid and also relatively more sensitive. Further, there is always a need for processes that use functional information (e.g., physical-chemical property information) to effectively extract useful information from sequence data being searched.

References

The following references (many of which are referred to above) to the extent that they provide exemplary procedural or other details supplementary to those set forth herein, are specifically incorporated herein by reference. Such references are not to be considered prior art to the present invention merely by such references being a part of this list.

Altschul, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389-3402.

Bairoch, et al. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Res.*, **28**, 45-48.

Benner, et al. (1994) Amino acid substitution during functionally constrained divergent evolution of protein sequences, *Protein Eng.*, **7**, 1323-1332.

Bowie, et al. (1991) A method to identify protein sequences that fold into a known three-dimensional structure, *Science*, **253**, 164-170.

Brenner, et al. (2000) The ASTRAL compendium for protein structure and sequence analysis, *Nucleic Acids Res.*, **28**, 254-256.

Chandonia, et al. (2002) ASTRAL compendium enhancements, *Nucleic Acids Res.*, **30**, 260-263.

Chothia, et al. (1986) The relation between the divergence of sequence and structure in proteins, *EMBO J.*, **5**, 823-826.

Dubchak, et al. (1999) Recognition of a protein fold in the context of the SCOP classification, *Proteins*, **35**, 401-407.

Eddy, S.R. (1998) Profile hidden Markov models, *Bioinformatics*, **14**, 755-763.

Falquet, et al. (2002) The PROSITE database, its status in 2002, *Nucleic Acids Res.*, **30**, 235-238.

Gough, et al. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments, *Nucleic Acids Res.*, **30**, 268-272.

Gribskov, et al. (1996) Identification of sequence patterns with profile analysis, *Methods Enzymol*, **266**, 198-212.

Henikoff, et al. (2000) Increased coverage of protein families with the Blocks Database servers, *Nucleic Acids Res.*, **28**, 228-230.

Henikoff, et al. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations, *Bioinformatics*, **15**, 471-479.

Henikoff, et al. (1994) Protein family classification based on searching a database of blocks, *Genomics*, **19**, 97-107.

Higgins, et al. (2000) Multiple sequence alignment, *Methods Mol. Biol.*, **143**, 1-18.

Holm, et al. (1996) Mapping the protein universe, *Science*, **273**, 595-602.

Kelley, et al. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM, *J. Mol. Biol.*, **299**, 499-520.

Kostich, et al. (2002) Human members of the eukaryotic protein kinase family, *Genome Biol.*, **3**, 43.

Kullback, et al. (1951) On information sufficiency, *Ann. Math. Stat.*, **22**, 79-86.

Lo Conte, et al. (2002) SCOP database in 2002: refinements accommodate structural genomics, *Nucleic Acids Res.*, **30**, 264-267.

Marcotte, et al. (1999) A combined algorithm for genome-wide prediction of protein function, *Nature*, **402**, 83-86.

Marcotte, E.M. (2000) Computational genetics: finding protein function by nonhomology methods, *Curr. Opin. Struct. Biol.*, **10**, 359-365.

Martelli, et al. (2002) A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins, *Bioinformatics*, **18**, S46-53.

Mehta, et al. (1999) Recognizing very distant sequence relationships among proteins by family profile analysis, *Proteins*, **35**, 387-400.

Nagano, et al. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions, *J. Mol. Biol.*, **321**, 741-765.

Norin, et al. (2002) Structural proteomics: developments in structure-to-function predictions, *Trends Biotechnol.*, **20**, 79-84.

Overbeek, et al. (1999) The use of gene clusters to infer functional coupling, *Proc Natl Acad Sci USA*, **96**, 2896-2901.

Rigoutsos, et al. (2002) Dictionary-driven protein annotation, *Nucleic Acids Res.*, **30**, 3901-3916.

Rison, et al. (2000) Comparison of functional annotation schemes for genomes, *Funct. Integr. Genomics*, **1**, 56-69.

Rychlewski, et al. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information, *Protein Sci.*, **9**, 232-241.

Schaffer, et al. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST- constructed position-specific score matrices, *Bioinformatics*, **16**, 488-189.

Schaffer, et al. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, *Nucleic Acids Res.*, **29**, 2994-3005.

Schein, et al. (2002) Total sequence decomposition distinguishes functional modules, "molegos" in apurinic/apyrimidinic endonucleases, *BMC Bioinformatics*, **3**, 37.

Urushihara, H. (2002) Functional genomics of the social amoebae, Dictyostelium

discoideum, *Mol. Cell*, **13**, 1-4.

Venkatarajan, et al. (2001) New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties, *J. Mol. Model.*, **7**, 445-453.

Waterston, et al. (2002) On the sequencing of the human genome, *Proc. Natl. Acad. Sci. USA*, **99**, 3712-3716.

Yona, et al. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory, *J. Mol. Biol.*, **315**, 1257-1275.

Zhu, et al. (2000) MASIA: recognition of common patterns and properties in multiple aligned protein sequences, *Bioinformatics*, **16**, 950-951.

Summary of the Invention

Generally, the present invention includes one or more of several processes or programs (or systems including such programs). For example, the present invention may be considered to include multiple (e.g., in one embodiment, three) processes or programs that may be implemented alone or in combination. Further, the output of one process may be used as an input by another program described herein or any other program that may operate on the input, the input to the process described herein may be receive from an output of another process, or the multiple processes may be used in any other effective combination.

For example, where the present invention is considered to include three processes or programs, a first program or process is operable to generate physical-chemical property (PCP) motifs from aligned sequences (e.g., an input multiple sequence alignment that can be obtained by various known methods).

A second program or process, for example, uses a matrix of values that define the PCP motifs, calculated by the first program or process, to search, in an automatic fashion, for related segments of protein sequences in a sequence database. For example, a positional scoring function based on related standard deviation and relative entropy values determined for positions of the initial multiple sequence alignment may be used. Thereafter, for example, the second program or process may select, for each protein in the sequence database, a highest scoring segment (e.g., such data may be stored for use in further processes or programs).

A third program or process may, for example, be used to analyze the data from the second process or program and determine which proteins (e.g., a ranking of such proteins) most resemble (e.g., in terms of PCP characteristics) the original aligned sequences (e.g., a family of related sequences). For example, the third program may use a Bayesian scoring function to perform such ranking of the proteins.

Further, additional related processes may incorporate the use of structural similarity algorithms. For example, after the above three programs detect significant homologues according to purely sequence criteria, and when structures are available, a process may calculate the segmental root-mean-square deviation (RMSD) between structures associated with the PCP query motif of the aligned sequences and those of the detected sequence segments of the proteins identified in the searched sequence database. The proteins may then be ranked according to overall structural similarity, or according to two criteria, similarity in physical-chemical properties and their overall structural similarity.

In one method of the present invention for use in sequence data analysis, the method includes providing a multiple sequence alignment of a plurality of sequences, wherein the multiple sequence alignment includes a column of aligned amino acids and/or gaps for each horizontal position of the multiple sequence alignment. Further, the method includes providing a plurality of numerical physical-chemical property (PCP) descriptors for each amino acid based on a plurality of physical-chemical properties thereof. Each of the plurality of numerical PCP descriptors corresponds to one of "N" eigenvectors used in defining the amino acids in terms of physical-chemical properties. Each amino acid in the multiple sequence alignment is quantitatively described in terms of the plurality of PCP descriptors as a series of "N" eigenvectors resulting in "N" PCP described sequence alignments, wherein each PCP described sequence alignment corresponds to and is defined with numerical PCP descriptors which correspond to one of the "N" eigenvectors, and further wherein each PCP described sequence alignment includes a plurality of columns corresponding to the columns of the multiple sequence alignment. Each of the PCP described sequence alignments is analyzed, on a column by column basis, to generate conservation property data for each column. The conservation property data for each column includes an average value for the numerical PCP descriptors in the column and a standard deviation associated with the average value, and a relative entropy value for the column. The conservation property data

for each of the PCP described sequence alignments is analyzed to detect consecutive horizontal positions of the multiple sequence alignment where the physical-chemical properties are conserved based on the relative entropy determined for each column. Thereafter, the method further defines one or more PCP motifs in the multiple sequence alignment based at least on the detection of consecutive horizontal positions of the multiple sequence alignment where the physical-chemical properties are conserved according to at least one eigenvector.

In one embodiment of the method, analyzing the conservation property data for each of the PCP described sequence alignments includes analyzing the conservation property data for each of the PCP described sequence alignments to detect consecutive horizontal positions where the relative entropy satisfies a predetermined limit.

In another embodiment of the method, defining one or more PCP motifs in the multiple sequence alignment further includes using user specified gap and minimum length limits to define the one or more PCP motifs, wherein each PCP motif comprises a plurality of consecutive horizontal positions in the multiple sequence alignment.

Yet further, in another embodiment of the method, the method includes using the one or more PCP motifs to search a sequence database for related sequence segments having PCP characteristics similar to one or more of the PCP motifs (e.g., defining each of the PCP motifs as a series of PCP motif profile matrices, wherein each PCP motif profile matrix of the series corresponds to one of the "N" eigenvectors, and further wherein values for each PCP motif profile matrix comprise an average value of the numerical PCP descriptors in the column at each horizontal position of the PCP motif and a standard deviation associated with the average value, and a relative entropy value for each horizontal position of the PCP motif).

In another embodiment, using the one or more PCP motifs to search a sequence database for related sequence segments includes converting each of one or more sequences of the sequence database to a searchable form using the numerical PCP descriptors, using a positional scoring function to match values of the series of PCP motif profile matrices defined for each PCP motif to segments of each of the searchable matrices resulting in scored segments, and selecting at least one scored segment for each of the searchable matrices as being a best match to each PCP motif based on results of the positional scoring function. Thereafter, the selected scored segments may be used for ranking the plurality of proteins of

the database according to which protein has PCP characteristics that are the closest to the plurality of sequences used to provide the multiple sequence alignment, for example, based on application of a Bayesian scoring function and/or based on structural similarity.

A computer program for use in conjunction with a processing apparatus to analyze sequence data is also provided. The computer program is operable when used with the processing apparatus to recognize a multiple sequence alignment of a plurality of sequences, wherein the multiple sequence alignment comprises a column of aligned amino acids and/or gaps for each horizontal position of the multiple sequence alignment. Further, the program is operable to recognize a plurality of numerical physical-chemical property (PCP) descriptors for each amino acid based on a plurality of physical-chemical properties thereof, wherein each of the plurality of numerical PCP descriptors corresponds to one of "N" eigenvectors used in defining the amino acids in terms of physical-chemical properties. Each amino acid in the multiple sequence alignment can be quantitatively described using the program in terms of the plurality of PCP descriptors as a series of "N" eigenvectors resulting in "N" PCP described sequence alignments. Each PCP described sequence alignment corresponds to and is defined with numerical PCP descriptors which correspond to one of the "N" eigenvectors. Further, each PCP described sequence alignment comprises a plurality of columns corresponding to the columns of the multiple sequence alignment. The computer apparatus is further operable to analyze each of the PCP described sequence alignments, on a column by column basis, to generate conservation property data for each column. The conservation property data for each column includes an average value for the numerical PCP descriptors in the column and a standard deviation associated with the average value, and a relative entropy value for the column. With further use of the program, the conservation property data is analyzed for each of the PCP described sequence alignments to detect consecutive horizontal positions of the multiple sequence alignment where the physical-chemical properties are conserved based on the relative entropy determined for each column and one or more PCP motifs in the multiple sequence alignment are defined based at least on the detection of consecutive horizontal positions of the multiple sequence alignment where the physical-chemical properties are conserved according to at least one eigenvector.

The computer program is further operable to carry out the various process steps described in one or more of the embodiments described herein.

The above summary of the present invention is not intended to describe each embodiment or every implementation of the present invention. Advantages, together with a more complete understanding of the invention, will become apparent and appreciated by referring to the following detailed description and claims taken in conjunction with the accompanying drawings.

Brief Description of the Drawings

Figure 1 shows a general block diagram of a general illustrative data processing system for use in analyzing a sequence database according to the present invention.

Figure 2 shows a general block diagram of a general illustrative sequence data analysis process according to the present invention.

Figure 3 shows a more detailed block diagram of one illustrative embodiment of a process for providing multiple sequence alignments as generally illustrated in the process of Figure 2.

Figure 4 shows a more detailed block diagram of one illustrative embodiment of a process for generating physical-chemical property motifs as shown generally in the process of Figure 2.

Figure 5 shows yet another block diagram of one illustrative embodiment of a process for generating physical-chemical property motifs as generally shown in the process of Figure 2, and also shown in the more detailed process of Figure 4.

Figure 6 shows a more detailed block diagram of one illustrative embodiment of a data mining process as generally illustrated in the process of Figure 2.

Figure 7 shows a more detailed block diagram of an illustrative embodiment of a protein ranking process shown generally in the process of Figure 2.

Figure 8 shows yet another block diagram of another protein ranking process as shown generally in the process of Figure 2.

Figures 9A-9C show tables of information for use in describing one example of the use of an analysis process according to the present invention.

Figures 10-E1 to 10-E4 show illustrations of the distribution of the values for each vector for the naturally occurring amino acids for use in describing one example of the use of an analysis process, according to the present invention. The frequency of occurrence of each

amino acid in SWISSPROT (release 40) was used.

Figure 11 shows one illustration of a qualitative representation of a motif 2 (from Figure 9A) for use in describing one example of the use of an analysis process according to the present invention. The magnitude of an eigenvector is shown as a positive or negative symbol depending on the average value at that position in the multiple alignment. A '*' indicates the relative entropy is less than 1.25 and the position will not be scored.

Figures 12A-12L show illustrations of the distribution of the highest scoring window for each motif (from the list of PCP motifs for the APE1 family shown in Figure 9A) in a specific (dotted lines) list of the scores for the 42 APE sequences and in the 3635 sequences of the non-specific ASTRAL40 database (solid lines) for use in describing one example of the use of an analysis process according to the present invention.

Figure 13 shows one illustration of a structure-based FSSP/DALI alignment of DNase-I family sequences, with motifs defined for the APE family underlined, for use in describing one example of the use of an analysis process according to the present invention. Motif areas that are structurally equivalent (as defined by DALI) are boxed.

Figure 14 shows a dendrogram of 42 different apurinic/apyrimidinic endonucleases and exonucleases used for the multiple alignment of the APE family for use in describing one example of the use of an analysis process according to the present invention. Number(s) indicate motif(s) that did not receive a significantly high score by subjecting the individual sequence of the protein to the motif search.

Detailed Description of the Embodiments

The present invention shall generally be described with reference to Figures 1 and 2. A more detailed description of one or more further embodiments of the present invention shall be described with reference to Figures 3-8. Thereafter, an example using a data analysis process according to the present invention shall be described with further reference to Figures 9-14.

Figure 1 shows a data analysis system 10 including a processing apparatus 12 along with associated data storage 14. Data storage 14 allows for access to processing programs or routines 16 and one or more other types of data 20 that may be employed to carry out the illustrative sequence data analysis method 30 as shown generally in the block diagram of

Figure 2.

For example, processing programs or routines 16 may include programs or routines for performing multiple alignment of sequences, physical-chemical property (PCP) based motif generation, data mining, or any other processing required to implement one or more embodiments of the present invention as described herein. Data 20 may include, for example, a sequence database, structural data associated with sequence information, results from one or more processing programs or routines employed according to the present invention, or any other data that is necessary for carrying out the one or more processes described herein.

As used herein, a protein sequence, also referred to herein solely as a sequence, refers to an order of the amino acids (each amino acid represented by a one letter code) in a protein (i.e., a protein molecule).

As used herein, a sequence database refers to any list of one or more sequences in a one letter code form that is in a format readable to a program or processing system. Examples of sequence databases include, for example, SwissProt, RefSeq, PIR, PRF, and PDB or user assembled specific lists of sequences. Further, whenever reference is made to a single sequence database, such a reference also includes a plurality of sequence databases as well. A sequence database also includes translations of nucleotide sequences in, for example, SwissProt (e.g., TrEMBL), GenBank and RefSeq. Such translations can include translations of annotated protein coding regions, and 6 frame translations, i.e., 3 reading frames from the forward nucleotide strand and 3 reading frames from the complementary nucleotide strand.

As used herein, a multiple sequence alignment refers to an alignment of a plurality of sequences with gaps inserted in the sequences to optimize alignment (e.g., to optimize alignment of residues with common structural positions in the same column of the multiple alignment). The alignment is preferably in a format readable to a program or processing system. As such, the multiple sequence alignment includes columns of aligned amino acids and/or gaps for each of a plurality of horizontal positions of the multiple sequence alignment (see, for example, the columns for each horizontal position in the alignment shown in Figure 13).

As used herein, a sequence segment, also referred to herein solely as a segment, refers to a portion of a sequence (i.e., sometimes also referred to as a sequence fragment).

As used herein, a PCP motif refers to a plurality of columns of the multiple sequence alignment at consecutive horizontal positions thereof determined as described herein based on at least the conservation of PCP characteristics.

In one or more embodiments of the present invention, the data analysis system 10 may be implemented using one or more computer programs executed on programmable computers, such as computers that include, for example, processing capabilities, data storage (e.g., volatile or non-volatile memory and/or storage elements), input devices, and output devices. Program code and/or logic described herein is applied to input data to perform functionality described herein and generate desired output information. The output information may be applied as input to one or more other devices and/or processes as described herein or as would be applied in a known fashion.

The program used to implement the present invention may be provided using any programmable language, e.g., a high level procedural and/or object orientated programming language, that is suitable for communicating with a computer system. Any such programs may, for example, be stored on any suitable device, e.g., a storage media, readable by a general or special purpose program, computer or a processor apparatus for configuring and operating the computer when the suitable device is read for performing the procedures described herein. In other words, at least in one embodiment, the system 10 may be implemented using a computer readable storage medium, configured with a computer program, where the storage medium so configured causes the computer to operate in a specific and predefined manner to perform functions described herein.

Likewise, the data analysis system 10 may be configured at a remote site (e.g., an application server) that allows access by one or more users via a remote computer apparatus (e.g., via a web browser), and allows a user to employ the functionality according to the present invention (e.g., user accesses a graphical user interface associated with the program to search a sequence data base, such as a public database).

The processing apparatus 12, may be, for example, any fixed or mobile computer system (e.g., a personal computer or mini computer). The exact configuration of the computing apparatus is not limiting and essentially any device capable of providing suitable computing capabilities may be used according to the present invention. Further, various peripheral devices, such as a computer display, mouse, keyboard, memory, printer, etc., are

contemplated to be used in combination with processing apparatus 12 of the data analysis system 14.

In view of the above, it will be readily apparent that the functionality as described in one or more embodiments according to the present invention may be implemented in any manner as would be known to one skilled in the art. As such, the computer language, the computer system, or any other software/hardware which is to be used to implement the present invention shall not be limiting on the scope of the processes or programs (e.g., the functionality provided by such processes or programs) described herein.

One will recognize that a graphical user interface, is used in conjunction with the embodiments described herein. The user interface may provide various features allowing for user input thereto, change of input, importation or exportation of files, or any other features that may be generally suitable for use with the processes described herein. For example, the user interface may allow default values to be used or may require entry of certain values, limits or other pertinent information.

Figure 2 shows a general block diagram of the illustrative generalized sequence data analysis method 30 according to the present invention. One will recognize that one or more of the blocks of functionality described therein may be carried out using one or more programs or routines. Further, certain of the functional blocks may be optional according to one more embodiments of the present invention.

Generally, the sequence data analysis method 30 includes providing the multiple sequence alignment (e.g., providing a multiple sequence alignment for a family of related sequences (block 32)). PCP motifs, e.g., PCP signatures, are then generated for the multiple sequence alignment (block 34) using multidimensional numerical PCP descriptors provided for each amino acid based on the physical-chemical properties of each amino acid. Such PCP motifs may then be used to mine a sequence database to locate related sequence segments (block 36) in a sequence database (block 38).

For example, such data mining may use a positional scoring function to detect related sequence segments of protein sequences of the sequence database. Thereafter, with use of information resulting from the data mining (block 36), one or more processes may be employed to rank the proteins associated with such related segments according to the proteins' resemblance to the PCP characteristics of the family of sequences used to generate

the multiple sequence alignment (block 40).

In general, and as will be further described herein, the present invention uses all the information in a multiple sequence alignment to identify common properties of the family of proteins used to generate the multiple sequence alignment. Thereafter, a user may use such identified properties to identify other proteins (e.g., from those proteins defined in a sequence database) that would fit into this family of proteins used to provide the multiple sequence alignment.

Further, and generally, the present invention is based on the conservation of physical-chemical properties of amino acids. The sequence data analysis method 30 provides a procedure in which each amino acid has been represented as a vector in a high-dimensional space determined by a plurality of different physical-chemical properties (e.g., a space determined by 237 different physical-chemical properties). By applying multidimensional scaling, the vector in high-dimensional space may be reduced into N-dimensions without losing information. For example, and as which will be described herein, in one embodiment, multidimensional scaling can reduce the vector in high-dimensional space into five dimensions without losing information. The properties of each amino acid in the reduced space can thus be described by five quantitative descriptors. Conservation of the physical-chemical properties may be reflected by the conservation of these descriptors.

Using a multiple sequence alignment of a plurality of sequences, the motif generating algorithm (block 34) takes the multiple alignment (block 32) and scores for conservation of the properties at each horizontal position of the multiple sequence alignment. For each position, the distribution of the values of the descriptors compared to natural distribution thereof can be expressed by the relative entropy. The magnitude of conservation can be measured as an average of the vectors at the horizontal position (e.g., an average of PCP descriptors for the amino acids in the column at the horizontal position). Using empirical rules, the present invention determines PCP motifs for the multiple sequence alignment. Each horizontal position in the PCP motif has a profile that includes average values for the descriptors and associated standard deviations, and further includes the relative entropies for the descriptors.

A positional scoring function may then be used to score the fit of a PCP profile matrix (e.g., the numerical and probabilistic expression of a PCP motif) to a given sequence in a

sliding window to detect segments in a given sequence as data mining of a searchable sequence database is performed to locate related sequence segments (block 36). Thereafter, one or more other scoring functions (e.g., Bayesian scoring function) may be used to determine which of a plurality of proteins of the searchable sequence database most resembles the family of proteins used to generate the multiple sequence alignment from which the PCP motifs are determined.

In other words, conservation in terms of these physical-chemical property eigenvectors in the multi-sequence alignment assists in identifying subtle signatures even among highly diverged protein family members. Such PCP motifs that are conserved in the physical-chemical properties in aligned protein sequences are important for biological function. The PCP motifs generated based on such conservation of the physical-chemical properties can be used to identify functionally related proteins in a sequence database which may have low sequence similarity. Sequence patterns based on conservation within the amino acid alphabet, as reflected in statistically derived substitution matrices, might not be sensitive enough to detect such subtle conservation of physical-chemical properties.

The superiority of the novel method according to the present invention when compared to current methods for determining sequence similarity is evidenced by identifying distant homologues of the human DNA repair enzyme, apurinic/apyrimidinic endonuclease 1 (APE1) (see the Example provided herein). All the commonly used methods for genome sequence searching rely on similar, statistically derived, scoring matrices. Frequently, the same scoring matrix is used to search for related sequences (for example, with BLAST), to prepare a multiple alignment of the protein sequences, to analyze sequence conservation, and finally, to locate distant relatives of the family according to motif conservation. The present invention, using PCP motifs, provides an alternative and an independent way to identify distantly related proteins based on sequence information.

For example, when the human APE1 sequence was used as a query for a PSI-BLAST search in the ASTRAL40 structural database with default parameters, neither known homologue of this enzyme in the database, bovine DNase-I or synaptojanin, a member of the Inositol 5'-polyphosphate phosphatase (IPP) family, was revealed. A PCP motif search according to the present invention showed, as the highest scoring proteins, all members of the DNase-I like SCOP-superfamily of APE in that database demonstrating that the present

invention can find non-trivial relationships between distantly related members within superfamilies. Other high scoring proteins were from different SCOP classifications but shared functions with the APE/DNase-I/IPP superfamily, including phosphatase activity and/or metal ion binding.

Further, the present invention may be used to identify functional areas in related proteins with different overall activities. In other words, besides the demonstrated capability of identifying distantly related proteins (or as a fold recognition tool), detecting PCP signatures or PCP motifs of a protein family is also helpful in identifying critical areas that are important for its function. Such PCP motifs can be further studied instead of the whole protein and these motifs can indicate target areas for developing drugs. Further, at least in one embodiment, by adding a structural definition to the sequence information allows a more focused detection system for determining relatedness of proteins. This use of "molegos," which refers to structurally related sequence motifs, may provide another way to identify functionally important areas of a protein. For example, molegos, and use thereof, are described in Schein, et al., Total sequence decomposition distinguishes functional modules, "molegos" in apurinic/apyrimidinic endonucleases, *BMC Bioinformatics*, **3**, 37 (2002), which is entirely incorporated herein.

The present invention uses specific mathematical tools which have been derived to locate PCP motifs in aligned protein sequences and to correlate sequences with structure and function. The present invention relies directly on the physical-chemical properties of amino acids and is substantially faster, relative to other methods, such as those based on neural networks or fold recognition tools. The present invention automatically identifies PCP motifs that can be used to detect common elements in proteins that share no global sequence identity.

Other methods that define motifs based on expert knowledge (e.g., such as those described in Falquet et al., The PROSITE database, its status in 2002, *Nucleic Acids Res*, 2002, 30:235-238; and Truong et al., Identification and characterization of subfamily-specific signatures in a large protein superfamily by a hidden Markov model approach, *BMC Bioinformatics*, 2002, 3(1):1) or groups of similar amino acids have just recently been described (e.g., such as in Ben-Hur et al., Remote homology detection: a motif based approach, *Bioinformatics*, 2003, 19(Suppl. 1):i26-i33) but are significantly different from the

method according to the present invention.

Databases that have been assembled with expert user knowledge, such as the PROSITE database (http://us.expasy.org/prosite/), which is a summary of data that may occur in other databases as well, is a list of motifs that have been defined as specific for a protein family. The syntax of the motif data storage system is limited to those specifically defined by the expert defining the motifs. The motifs are not automatically derived and are described in alphabetic fashion, and not according to physical chemical properties as in the method described according to the present invention.

The motifs defined by the E-motif program (http://fold.stanford.edu/emotif/emotif-maker.html) are automatically derived but are not defined according to aggregate physical chemical properties in a systematic fashion such as that described according to the present invention. The approach to sequence database searching is different and less systematic than that according to the present invention. Neither the E-motif nor the PROSITE approach seeks to link structural data to further define motifs in the form of molegos, conserved units in protein families that are conserved in both structure and sequence.

Further, for databases, the motifs are defined to discriminate one protein family from another. At least in one embodiment of the present invention, the novelty of the approach according to the present invention is that the PCP motifs are used to completely describe the functions and structural characteristics of a given protein family. As such, motifs (and molegos) can be defined as specific for a protein with a given function, or as general (or generic) to many different types of proteins that may share a common function. Hence, the method may be referred to as "total sequence decomposition" and, at least in one embodiment, is distinguished by the programs according to the present invention being prepared to perform such decomposition automatically so as to be applicable to genomic database screening.

One or more various embodiments of the illustrative sequence data analysis method 30, shown generally in Figure 2, shall be described with reference to Figures 3-8. The provision of the multiple sequence alignment (block 32), shown generally in Figure 2, may be provided by any suitable alignment tool. For example, such alignment tools include BLAST available on the internet through the National Center for Biotechnology Information (NCBI) (website .ncbi.nlm.nih.gov/BLAST), CLUSTALW available on the internet through the

European Bioinformatics Institute (EBI) (website ebi.ac.uk/clustalw/), or any other suitable multiple alignment tool.

One embodiment of providing a multiple sequence alignment 32 is shown in Figure 3. For example, diversely related sequences may be collected (block 50) by any suitable tool such as BLASTP available on the internet through the National Center for Biotechnology Information (NCBI), or any other suitable tool. In other words, for example, related sequences of a known protein family desired to be used for query of a sequence database may be collected. Thereafter, a multiple sequence alignment is generated (block 52) for the related sequences of the family such as with an alignment tool as described herein. Further, the multiple sequence alignment is provided for use by the PCP motif generation program (block 54).

For example, according to one embodiment of the present invention, the multiple sequence alignment includes a column of aligned amino acids and/or gaps for each of a plurality of horizontal positions of the multiple alignment. In other words, the alignment may include q sequences and be defined using s columns. In such a case, the alignment takes the form of a matrix of amino acid codes and gaps which can be expressed as *sxq*. A file representative of the multiple sequence alignment may then be provided for use in generating PCP motifs (block 34) of the sequence data analysis method 30.

The PCP motif generation process 34 is further described in detail with reference to Figures 4 and 5. As shown in the block diagram of Figure 4, the PCP motif generation process 34 is provided with the multiple sequence alignment (block 60), such as that provided as described with reference to Figure 3. Further, multidimensional numerical PCP descriptors for each amino acid based on the physical-chemical properties of each amino acid are provided (block 62) so that the multiple sequence alignment can be quantitatively described in terms of the PCP descriptors as a series of "N" eigenvectors (block 64).

The multidimensional numerical PCP descriptors for each amino acid (block 62) may be determined as described in the article by Venkatarajan and Braun, entitled "New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties," *J Mol. Model.* (2001) 7:445-453, which is entirely incorporated herein by reference. As described therein, quantitative descriptors for the 20 naturally occurring amino acids based on multidimensional scaling of 237 physical-chemical

properties are provided. In one embodiment, a five-dimensional property space can be constructed such that the amino acids are in a similar spatial distribution to that in the original high-dimensional property space. Properties that correlate well with the five major components were hydrophobicity, size, the relative tendency for each of the amino acids to occur in α-helices, the number of degenerate triplet codons that could encode the amino acid and the frequency of occurrence of amino acid residues in β-strands.

The quantitative descriptors summarize information about physical-chemical properties that are useful in identifying protein homologues on the basis of PCP-based motifs. Multidimensional scaling of some or all of 237 physical-chemical properties can be used to derive quantitative descriptors for all 20 naturally occurring amino acids, and the main variations of all properties for the 20 amino acids can be reduced, for example, through five quantitative descriptors. Multidimensional scaling is a general classification approach to reconstructing the geometrical configuration of large point sets in lower dimensions. While the physical meaning of each descriptor can be correlated with individual properties, the five descriptors cannot simply be replaced by five individual properties.

The quantitative descriptors described in Venkatarajan and Braun (2001), or as otherwise developed using the same or similar techniques, can be used to characterize PCP motifs in protein families through employment of the present invention.

In one particular embodiment of providing the quantitative descriptors, each property was normalized, where S is normalized property values, α is the index of the property and i stands for the amino acid. P is the property value, $\overline{P}_\alpha$ and $\sigma_{P\alpha}$ are the average and the standard deviation of property α.

Each amino acid i is represented as a vector $\underline{S}(i)$ in a 237-dimensional continuous space, where the components $S_\alpha(i)$ are the normalized property values. The scalar product Q_{ij} between two vectors $\underline{S}(i)$ and $\underline{S}(j)$, where j is another index for an amino acid, is given by

$$Q_{ij} = \underline{S}(i) \bullet \underline{S}(j)$$

$$Q_{ij} = \sum_{\alpha=1}^{237} S_\alpha(i) \cdot S_\alpha(j)$$

The positive symmetric 20x20 matrix Q consists of the scalar products of the property vectors $\underline{\mathbf{S}}(\boldsymbol{i})$ and $\underline{\mathbf{S}}(\boldsymbol{j})$, where i=1...20 and j=1...20.

Eigenvectors E and eigenvalues λ of the matrix Q can be computed using the JACOBI

and EIGSRT subroutines provided in Numerical Recipes (Press, W.H., Teukolsky, S.A., Vettering, W.T., Flannery, B.P., 1999, "Numerical recipes in C", Cambridge University Press, New York).

$$Q \cdot E = \lambda E$$

As Q is of order 20, there will be 20 eigenvectors and eigenvalues λ and the smallest eigenvalue λ_{20} is equal to 0 due to normalization of the properties. The subroutine JACOBI implements a Jacobian transformation of a symmetric matrix and returns the eigenvectors and values of the Q matrix. The eigenvalues and their corresponding eigenvectors can be indexed in decreasing order of the eigenvalues.

If μ represents the index of eigenvalue and eigenvector, then the elements of the Q matrix can be equated to eigenvalues and eigenvectors as:

$$Q_{ij} = \sum_{\mu=1}^{20} \lambda_\mu E_i^\mu \cdot E_j^\mu$$

The first five significant eigenvalues were selected for the representation of amino acids, thus Q_{ij} can be written as:

$$Q_{ij} \approx \sum_{\mu=1}^{5} \lambda_\mu E_i^\mu \cdot E_j^\mu$$

Each amino acid can be represented as a vector in the five-dimensional Euclidean space (a.k.a. Eigen sub-space) with each dimension perpendicular to each other. The coordinates of the *i*th amino acid can be written as:

$$\sqrt{\lambda}_{\mu=1} E_i^{\mu=1}, \sqrt{\lambda}_{\mu=2} E_i^{\mu=2}, \sqrt{\lambda}_{\mu=3} E_i^{\mu=3}, \sqrt{\lambda}_{\mu=4} E_i^{\mu=4}, \sqrt{\lambda}_{\mu=5} E_i^{\mu=5}$$

The distance between the *i*th and *j*th amino acids is given by:

$$d_{ij} = \sqrt{\sum_{\mu=1}^{5} \left(\sqrt{\lambda}_\mu E_i^\mu - \sqrt{\lambda}_\mu E_j^\mu \right)^2}$$

Distances computed between amino acids in the five-dimensional Eigen sub-space constitute the property distance matrix (PDM). Small distance values between two amino acids indicate they are similar in all of their 237 physical-chemical properties.

The distribution of the eigenvalues of the Q matrix, containing the scalar products between all pairs of the 237-dimensional amino acid vectors, rapidly decreases from the largest value λ_1=1962 to λ_{19}=16. The rapid decrease of the eigenvalues derived from the 237 physical-chemical properties suggests that the number of properties can be reduced while

retaining approximately the same distribution of amino acids in the property space. The eigenvalues rapidly decrease within the five largest eigenvalues. As such, the first five eigenvalues and eigenvectors can be used to calculate five-dimensional numerical descriptors of the amino acids. The five numerical descriptors for each amino acid are given in Table 1.

Table 1

Components E1 to E5 of 237 physical-chemical properties for each amino acid

Eigenvalue (λ)	Eigenvector[a]				
	E1	E2	E3	E4	E5
	1961.504	788.200	539.776	276.624	244.106
A	0.008	0.134	-0.475	-0.039	0.181
R	0.171	-0.361	0.107	-0.258	-0.364
N	0.255	0.038	0.117	0.118	-0.055
D	0.303	-0.057	-0.014	0.225	0.156
C	-0.132	0.174	0.070	0.565	-0.374
Q	0.149	-0.184	-0.030	0.035	-0.112
E	0.221	-0.280	-0.315	0.157	0.303
G	0.218	0.562	-0.024	0.018	0.106
H	0.023	-0.177	0.041	0.280	-0.021
I	-0.353	0.071	-0.088	-0.195	-0.107
L	-0.267	0.018	-0.265	-0.274	0.206
K	0.243	-0.339	-0.044	-0.325	-0.027
M	-0.239	-0.141	-0.155	0.321	0.077
F	-0.329	-0.023	0.072	-0.002	0.208
P	0.173	0.286	0.407	-0.215	0.384
S	0.199	0.238	-0.015	-0.068	-0.196
T	0.068	0.147	-0.015	-0.132	-0.274
W	-0.296	-0.186	0.389	0.083	0.297
Y	-0.141	-0.057	0.425	-0.096	-0.091
V	-0.274	0.136	-0.187	-0.196	-0.299

[a] The numerical descriptors for each amino acid *i* are calculated by $\sqrt{\lambda}_{\mu}E_i^{\mu}$ for the five eigenvectors $E^{\mu}, \mu = 1..5$.

Further, the quantitative descriptors represent a precise spatial relation of all amino acids with respect to many physical-chemical properties.

Although the above embodiment describes the use of five-dimensional descriptors, it will be recognized that any number of N-dimensional descriptors may be used. In other words, the quantitative descriptors may be based on more or less than five eigenvectors for the amino acids. However, in one preferred embodiment, five-dimensional quantitative descriptors for the physical-chemical properties for each amino acid are provided. In other

words, the numerical PCP descriptors for each amino acid include a numerical descriptor for each of a plurality of N eigenvectors (e.g., five eigenvectors where five-dimensional descriptors are used).

Further, with reference to Figure 4, after the multiple sequence alignment is quantitatively described in terms of the PCP descriptors as a series of N eigenvectors (block 64), the quantitatively PCP described multiple sequence alignment is analyzed based on the conservation of the PCP properties at each horizontal position of the alignment (block 66). As shown generally by block 66, the quantitatively PCP described sequence alignment is analyzed, on a column-by-column basis (i.e., for each horizontal position of the alignment), as a function of the conservation of the PCP descriptors in each column in the series of eigenvectors. The conservation is measured using standard deviation of the values of the PCP descriptors for each column analyzed and the related relative entropy for the analyzed column.

Generally, PCP motifs may be defined for the multiple sequence alignment at consecutive positions of the multiple sequence alignment where significant conservation exists, for example, where the relative entropy exceeds a predetermined limit at consecutive horizontal positions of quantitatively PCP described alignment (block 68). Once the PCP motifs have been defined, a series of PCP motif profile matrices may be generated for each PCP motif (e.g., a multi-dimensional matrix) (block 70), such as for use in data mining of a sequence database (block 36, Figure 2).

In other words, when considering a five-dimensional space configuration, the quantitative descriptors E^1 to E^5 for amino acid properties and their physical interpretation are deduced from a comprehensive list of 237 physical-chemical properties; the five components E^1 to E^5 will in general be differently conserved in each column (k) of the aligned sequences of a protein family. The conservation within a column of the multiple alignment can be measured by standard deviations σ_k^i of the numerical values at that position for eigenvectors E^1 to E^5 (where i varies from 1-5) and by the relative entropy $\Re_k^i$, also referred to as Kullback-Leibler distance (Kullback, S. and Leibler, R.A., 1951, "On information sufficiency", *Annals Math. Stat.*, **22**, 79-86), that describes the relative conservation of the amino acids in the column as compared to that expected for a random distribution. The component index *i* varies from 1 to 5, where n=5-dimensional space is used. The quantities

σ_k^i and $\Re_k^i$ are calculated across the residues in each column k in the PCP descriptor described alignment of the protein family of sequences (block 66).

Five equally spaced bins characterize the distributions of the E1 to E5 values for each of the components. The difference between the observed distribution for the component E^i at column k and the background distribution can be calculated by the relative entropy $\Re_k^i$:

$$\Re_k^i = \sum_{b=1}^{5} Q(X^b) \log_2 \left(\frac{Q(X^b)}{P(X^b)} \right)$$

$Q(X^b)$ is the observed fraction of the component i in the bin b and $P(X^b)$ is the corresponding background frequency. For significantly conserved properties, the distributions of the component values E^1 to E^5 are narrower than the background distributions derived from the 'a priori' occurrences of amino acids; i.e., low standard deviations and high relative entropy values would be characteristics of such significantly conserved properties. If the distributions of the observed frequencies of the components are equal to that of the background frequencies, then $\Re$ will be zero, otherwise it will be positive. High relative entropy values indicate a significant difference between the observed frequencies of distributions in a column and the 'a priori' background distribution.

PCP motifs may be defined as contiguous horizontal positions in a multiple alignment where residues are significantly conserved according to one of the principal components E^i, that is those columns k where the relative entropy $\Re_k^i$ of at least one component E^i is above a $\Re$-cutoff value (block 68). A minimum length cutoff (L-cutoff) is provided by the user to define PCP motifs of sufficient length. The user can also specify a G-cutoff parameter (i.e., a gap limit) to define the maximum number of insignificant positions between two significant positions in a PCP motif. Default values of the parameters $\Re$-cutoff, L-cutoff and G-cutoff can be determined empirically.

Each PCP motif identified in the PCP described alignment may be quantitatively expressed as a profile including the average values, standard deviations, and the relative entropies for each eigenvector E1-E5 at each horizontal position of the alignment (i.e., corresponding to each column in the alignment). Such values are used to generate a PCP motif profile matrix (block 70) for use in data mining of a sequence database.

Figure 5 provides a more detailed block diagram of the PCP motif generation process

34 according to one embodiment of the present invention. As shown therein, with provision of the multiple sequence alignment (block 60) and the provision of multidimensional numerical PCP descriptors (block 62), each amino acid of the multiple sequence alignment can be quantitatively described in terms of the PCP descriptors resulting in a plurality of distinct PCP described alignments (block 84). Each distinct PCP described alignment corresponds to and is defined with numerical PCP descriptors that correspond to one of the N eigenvectors. Further, each distinct PCP described alignment includes a plurality of columns of descriptor values (e.g., values that correspond to the appropriate eigenvector) corresponding to the columns in the multiple sequence alignment.

For example, the plurality of distinct PCP described alignments that represent the multiple sequence alignment can be considered to include N distinct *sxq* matrices, where *s* is the number of horizontal positions or columns of the alignment and *q* is the number of sequences. Each PCP described alignment or matrix includes descriptor values for the amino acids corresponding to the respective eigenvector of the series of eigenvectors to which the alignment corresponds (e.g., the gaps being left as zero). For example, a PCP described alignment corresponding to eigenvector E1 would include amino acid numerical descriptors corresponding to the eigenvector E1, a PCP described alignment corresponding to eigenvector E2 would include amino acid numerical descriptors defined for the E2 eigenvector, and so forth. In other words, the multiple sequence alignment is described quantitatively in terms of PCP descriptors for eigenvectors E1-E5 for a five-dimensional configuration. Such a series of PCP described alignments may be thought of as a multi-dimensional alignment with the third dimension being N, where N is the number of eigenvectors or dimensional space being used (e.g., 5-dimensional space); the other dimensions including s and *q*.

Each distinct PCP described alignment (e.g., corresponding to E1-E5 and including five alignment matrices in a five-dimensional configuration) is analyzed (block 86). Such analysis is done on a column-by-column basis for each horizontal position of each PCP described alignment to generate conservation property data corresponding thereto. In other words, each column of a PCP described alignment is analyzed. Such analysis results in values to be used in detecting conservation of the PCP characteristics. The values for each column include an average value of the PCP descriptors of each column and the associated

standard deviation along the column being analyzed, as well as a relative entropy for the column being analyzed (see description provided with reference to Figure 4 for algorithms to provide such values). Such values are determined for each horizontal position (i.e., *s*) in the PCP described alignment. Further, each PCP described alignment (e.g., five alignments in the 5-dimensional configuration) is likewise analyzed.

Thereafter, the conservation property data generated for each of the PCP described alignments (i.e., corresponding to each of the eigenvectors being considered) is scanned (e.g., the conservation property data generated for each column or horizontal position of the alignment is scanned) to detect consecutive horizontal positions of the alignment (i.e., columns of an alignment) where the relative entropy exceeds a predetermined limit (block 88). The predetermined limit is set by a user and determined as a function of the sequence variability of the protein family under consideration.

Based on the detected consecutive positions (block 88) and also user-specified gap and minimum length limits (block 92), PCP motifs are located in the multiple sequence alignment (block 90). The user-specified gap and minimum length limits (block 92) are user input values. In other words, as a result of the detected consecutive conserved positions and user-specified gap and minimum length limits, PCP motifs in the original multiple sequence alignment are defined. It is noted that such conservation may be detected with respect to any of the eigenvectors (e.g., represented by the respective PCP described alignments) resulting in a PCP motif related to such detected conservation. A list of the PCP motifs may then be provided in a desired form (e.g., a listing of the columns of the multiple sequence alignment identified). The list of such PCP motifs may be printed out or saved to a file.

The program may then generate a series of PCP motif profile matrices for each PCP motif (block 94). Each matrix of the series of PCP motif profile matrices corresponds to one of N eigenvectors. The program compiles values for each horizontal position (e.g., column) of the PCP motif to be used in a corresponding profile matrix. The values for each PCP motif profile matrix include the average value of the PCP descriptors for each horizontal position in the PCP motif (e.g., those amino acid descriptors, for the corresponding eigenvector, provided for each residue in each column of the PCP motif), and the standard deviation associated with the average value, as well as the relative entropy for the horizontal position in the motif (e.g., column of the PCP motif).

As such, at least in one embodiment, the series of PCP motif profile matrices could be expressed as a multi-dimensional matrix, *mxn*(p), where *n* is the number of eigenvectors and m is the number of horizontal positions (e.g., columns) in the PCP motif, and further wherein p are the number of values generated for the matrix (e.g., where p is 3 when the values include average value, standard deviation and relative entropy).

Further, in another embodiment, each of the series of PCP motif profile matrices may each include a first level that includes an average value of PCP descriptors for each horizontal position in the PCP motif for the corresponding eigenvector and the standard deviation associated therewith, and a second level that gives the relative entropy for each horizontal position (i.e., column) of the PCP motif for the corresponding eigenvector.

With further reference to Figure 2, after generation of the PCP motifs (block 34), such PCP motifs may be used for mining sequence data to locate related segments of protein sequences in a database having PCP characteristics similar to the PCP motifs (block 36). The series of PCP motif profile matrices (or multidimensional matrix) for each PCP motif can be used to perform such data mining.

Figure 6 provides one embodiment of a data mining process 36 that may be used according to the present invention. One will recognize that although a positional scoring function is utilized in this particular embodiment, that one or more other types of scoring functions, including other positional scoring functions, may be utilized to match PCP motif profile matrices to segments of sequence data being searched. Generally, any scoring function that relates the sequence windows of the database in a probabilistic fashion to the average values of the properties determined for the motifs (block 94) can be used. However, meaningful values can only be obtained if the absolute values are expressed as an expectation value according to the average, or random, distribution of amino acids.

As shown in the block diagram of Figure 6, in one embodiment of the data mining process 36, the process includes providing one or more PCP motif profile matrices (block 100) and also providing a sequence database to be searched (block 102). The process then uses the values of the PCP motif profile matrices to search, in an automatic fashion, for related segments of protein sequences in the sequence database.

In this embodiment of the data mining process 36, the process provides for conversion of the protein sequences of the database to a searchable form suitable for use with PCP motif

profile matrices (block 104). For example, whole sequences are converted to matrices using the numerical PCP descriptors (e.g., *mxn* matrices, where *n* is the number of eigenvectors and m is the length of the sequence). Further, for example, nucleotides (e.g., RNA, DNA) may be converted to amino acid sequences, and thereafter to a suitable and searchable form. The term sequence database includes the searching of nucleotides as well as amino acid lists).

Values of the PCP motif profile matrices are then matched to the converted sequence matrices using a positional scoring function (block 106). In one embodiment, for example, a Lorentzian based scoring scheme may be used to measure the quality of fit for a database sequence to a PCP motif profile matrix at position *k* for the component eigenvector *i*. The PCP motif profile matrix at a significant position *k* (defined by the relative-entropy cutoff) includes an average of component magnitudes $< E_k^i >$ and the standard deviation σ_k^i. If E^i is the magnitude of PCP component *i* observed in the query sequence, the *Z* value is calculated:

$$Z_k^i = \left(\frac{E^i - < E_k^i >}{W\sigma_k^i + \Phi} \right)$$

$$S_k^i = \left(\frac{1}{1 + Z_k^i * Z_k^i} \right)$$

Where *W* is the weight for standard deviation (e.g., set to 1.5 in one embodiment to perform calculations) and Φ is a small positive shift (e.g., set to 0.001 in one embodiment) added to the denominator to prevent overflow during calculation when σ_k^i is zero. The individual score for each component is then added for significantly conserved property components along the length of the motifs to obtain a window score S_w and a maximum possible score S_{max} calculated by adding 1 for every significant position:

$$S_w = \sum_{i,k} S_k^i$$

$$S_{\max} = \sum_{i,k} 1$$

The final fractional score for the window is:

$$S_{wfraction} = \frac{S_w}{S_{\max}}$$

Thereafter, at least in one embodiment, the process 36 selects, for each protein in the sequence database, the highest scoring window as being the best segment match for the series

of PCP motif profile matrices defining the PCP motif (block 108) for that particular protein. For example, the program may select, for each protein in sequence database being searched, the highest scoring window, and store it in a file or temporary buffer (e.g., a data file that may be used by one or more subsequent programs or processes). One will recognize that multiple high scoring segments may also be selected, and the present invention is not limited to just the highest scoring window.

As shown in block 110, a file is provided including values and/or sequences of the segments that are best matched to the PCP motif profile matrices. In other words, the file may include values and sequences of highest scoring segments for each protein in the searched database. Such matched areas of the proteins have similar physical-chemical properties to the average value for the sequence family used to generate the multiple sequence alignment but may include a different amino acid string than a consensus sequence generated from the multiple sequence alignment according to a conventional method.

For conditions or applications where a user only desires to find which areas of a protein are best matched to the list of PCP motifs, the program may provide the file and no other processing need be performed. However, optionally, such located segments that match the PCP motifs can further be used to rank the proteins of the database according to their similarity to the original family of protein sequences used to generate the multiple sequence alignment. For example, the data mined from the exemplary embodiment of Figure 6 may be provided for use in determining which proteins of the sequence database most resemble the original family of proteins used to provide the multiple sequence alignment (block 40). Several related protein ranking processes (e.g., selection processes) are described further herein with reference to Figures 7 and 8.

As shown in the block diagram of Figure 7, one embodiment of the related protein ranking process 40 includes providing the output of the positional scoring function process, including the scoring data for the highest scoring segments as described with reference to Figure 6 (block 120). In other words, for each of the proteins in the sequence database, the highest scoring window (or windows) for each of the PCP motifs is selected.

In one embodiment, an optional linking process is performed to link the data for each protein. For example, in one embodiment, the PCP motifs are individually matched, in consecutive order, and a final list of each highest scoring window in each sequence in the

database is stored. The lists are then linked for each protein after mining process is performed (block 122). As an alternative, each sequence may be individually evaluated and summarized, which may save storage for large sequence databases.

Thereafter, an overall similarity distance score for each protein of the sequence database being searched is calculated or otherwise determined, based on the score for each of the highest scoring segment of each protein and relative to what a random score would be for the segment (block 124).

In one embodiment, a Bayesian scoring function is used to determine the overall similarity distance for scoring the proteins. For example, a Bayesian method is used to decide if a given score S for a segment in an arbitrary searchable sequence is a sufficient match to a PCP motif, relative to what a random sequence would score. The conditional probability $P(X \in PCP \mid S)$ that the queried sequence X contains a PCP motif for a given score S is given by Bayes theorem:

$$P(X \in PCP|S) = \frac{P(S|X \in PCP) \bullet P(PCP)}{P(S)}$$

$P(S \mid X \in PCP)$ is the probability of finding a motif with score S in the sequence family used to generate the PCP motif. $P(S)$ is the probability of finding a PCP motif with similar score in all proteins in the database being searched, and $P(PCP)$ is the probability of finding sequences of the family in the searched database. The empirical distributions of scores in the sequence family and searched database are approximated as a Gaussian distribution:

$$P(S|X \in PCP) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(S-\bar{S}_{PCP})^2}$$

Here, $\bar{S}_{PCP}$ is the average score for the PCP motif and σ is the corresponding standard deviation. The probability of finding a PCP motif with similar score in the searched database (DB) is:

$$P(S) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(S-\bar{S}_{DB})^2}$$

Substituting the above two expressions into Bayes theorem, and simplifying by assuming that the standard deviation of scores for PCP and the searched database are similar, we obtain:

$$P(X \in PCP|S) = e^{\frac{\Delta\bar{S} \bullet (S-\bar{S})}{\sigma^2}} \bullet P(PCP)$$

Here, $\bar{S}$ is the average and $\Delta\bar{S}$ is the difference between average scores of a PCP motif in the searched database. The conditional probability that a PCP motif *m* was found in a searchable sequence *X* with a score less than or equal to an observed score S_m is given by:

$$P(X \in PCP | S \le S_m) = P(PCP) \bullet \int_0^{S_m} e^{\frac{\Delta\bar{S}m \bullet (S - \bar{S}_m)}{\sigma_m^2}} dS$$

$$P(X \in PCP | S \le S_m) = P(PCP) \bullet \left[\frac{\sigma_m^2}{\Delta\bar{S}_m} e^{\frac{-\Delta\bar{S}_m \bullet \bar{S}_m}{\sigma_m^2}} \right] \bullet \left[e^{\frac{\Delta\bar{S}_m \bullet Sm}{\sigma_m^2}} - 1 \right]$$

We compute a total sequence score S_x over all NP PCP motifs located in the family of sequences (e.g., the multiple sequence alignment) as:

$$S_X = \sum_{m=1}^{NP} \log_2 [P(X \in PCP | S \le S_m)]$$

$$S_X = \sum_{m=1}^{NP} \log_2 \left[\frac{\sigma_m^2}{\Delta\bar{S}_m} e^{\frac{-\Delta\bar{S}_m \bullet \bar{S}_m}{\sigma_m^2}} \right] + \sum_{m=1}^{NP} \log_2 \left[e^{\frac{\Delta\bar{S}_m \bullet S_m}{\sigma_m^2}} - 1 \right] + NP * \log_2 P(PCP)$$

The first and the third terms of the last equation are constant for a given database so the middle term is used for a final scoring and ranking of proteins.

As a result of the Bayesian scoring method, an overall PCP similarity score is provided for the proteins of the sequence data being searched and such scores may be ranked (block 126). With use of the ranked proteins in the searched database, according to their similarity distance score, an output file, including the proteins with the highest score(s) (block 128), can be provided to a user as requested. For example, a user may desire receipt of the top four highest scoring proteins to be listed at the user interface.

As shown in the block diagram of Figure 8, structural similarity may be used in combination with PCP similarity (such as described with reference to Figure 7), or alone, to determine which proteins of the searched sequence data most resemble the multiple sequence alignment. As shown in Figure 8, the protein ranking process 40, according to yet another embodiment, includes providing a file including structure for a plurality of segments that are best matched to the one or more PCP motif profile matrices (block 142). In addition, query structures related to the PCP motifs of the multiple sequence alignment are also provided (block 148), e.g., molegos. For example, 3D structure relating to one or more proteins of the

family of proteins used in providing the original multiple sequence alignment is provided for comparison to the best-matched segments.

The segmental root-mean-square deviations (RMSD) between the query structures and the structures of the best matched segments are calculated (block 146). In other words, at least in one embodiment, after one or more processes detect significant homologues in the sequence database according to purely PCP similarity (e.g., high scoring matches between the PCP motif and segments of proteins in the sequence database), and when 3D structures are available, such calculations may be made. The structure of a relevant comparison protein may be determined experimentally or by theoretical modeling.

As shown in block 150, the proteins of the searched sequence database for which the structural similarity is being compared may be ranked according to structural similarity as determined using the segmental RMSD. Further, scoring of the proteins of the database according to PCP similarity (block 154) may be provided such that the proteins of the searched sequence data may be ranked according to both structural similarity and PCP similarity (block 152) to set forth proteins which most resemble the original family of proteins used to provide the multiple sequence alignment.

Example-Use of present invention to determine distant homologues of APE1

The following provides one example of the use of a tool according to the present invention. The PCP software tool was used to generate PCP motifs from an alignment of proteins belonging to the DNA repair protein family APE and then search for proteins in the ASTRAL40 database containing similar sequences. The highest scoring proteins were in the DNase-I like SCOP-superfamily of APE, demonstrating that the process according to the present invention can find non-trivial relationships between distantly related members within superfamilies. Other high scoring proteins were from different SCOP classifications (Lo Conte, et al., 2002, SCOP database in 2002: refinements accommodate structural genomics, *Nucleic Acids Res.*, **30**, 264-267), but shared functions with the APE/DNase-I/IPP superfamily, including phosphatase activity and/or metal ion binding. Details of the structural and functional roles of the PCP motifs of the APE family are described in Schein, et al., 2002, Total sequence decomposition distinguishes functional modules, "melogos" in apurinic/apyrimidinic endonucleases, *BMC Bioinformatics*, **3**, 37.

PCP motifs were generated for the APE protein family. Homologues of human APE1

with E-values less than 0.001 were identified in the NCBI protein sequence database using the BLASTP search engine (Altschul, et al., 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389-3402). Sequences from 42 organisms ranging from prokaryotes to eukaryotes were selected (see Figure 14) after discarding hypothetical APE-like proteins. The taxonomic classification was used to avoid excessive redundancy. Sequences were aligned with CLUSTALW release 1.8 (Higgins, et al., 2000, Multiple sequence alignment, *Methods Mol. Biol.*, **143**, 1-18) using the GONNET similarity matrix (Benner, et al., 1994, Amino acid substitution during functionally constrained divergent evolution of protein sequences, *Protein Eng.*, **7**, 1323-1332), with an opening gap penalty of 10.0 and gap extension penalty of 0.2. The sequence alignment was used as input for the PCP generation program for motif identification, as shown and described with reference to Figure 5.

Each motif identified is quantitatively expressed as a profile, including the average values, standard deviations and the relative entropies for each vector E1-E5 at each position (column in the initial alignment) in the motif. This profile was used to search for similar motifs in the ASTRAL40 sequence database (Brenner, et al., 2000, The ASTRAL compendium for protein sequence and structure analysis, *Nucleic Acids Res.*, **28**, 254-256; Chandonia, et al., 2002, ASTRAL compendium enhancements, *Nucleic Acids Res.*, **30**, 260-263), which consists of representative sequences corresponding to the SEQRES record of PDB files and classified in SCOP (Lo Conte, et al., 2002, SCOP database in 2002: refinements accommodate structural genomics, *Nucleic Acids Res.*, **30**, 264-267). The pair wise identity among the 3635 sequences is less than 40%.

A Lorentzian based scoring system and the Bayesian scoring method described herein were utilized to provide a ranking of the protein sequences in the ASTRAL40 database.

The sensitivity of the present invention to find proteins related to the APEs in the ASTRAL40 database were compared with that of two versions of PSI-BLAST with default parameters (E-value of 0.005), one locally installed (v 2.2.1) and the other on the web at NCBI (Altschul, et al., 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389-3402; Schaffer, et al. , 2001, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, *Nucleic Acids Res.*, **29**, 2994-3005). To enhance the ability

of PSIBLAST to build a profile, the 42 sequences from the APE family, used in the construction of the motifs, were added to the 3635 sequences from the ASTRAL40 database. The human APE sequence was used as query and ran up to five iterations of PSI-BLAST. A search for APE related sequences in the BLOCKS database with the default search engine (Henikoff, et al., 2000, Increased coverage of protein families with the Blocks Database servers, *Nucleic Acids Res.*, **28**, 228-230; Henikoff, et al., 1999, Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations, *Bioinformatics*, **15**, 471-479; Henikoff, et al., 1994, Protein family classification based on searching a database of blocks, *Genomics*, **19**, 97-107) was also performed.

The 'a priori' distribution of the property components E^1 to E^5

A comparison was performed for the distribution of the 20 amino acids according to each of the five property components based on an 'a priori' distribution derived from their relative occurrence in the SWISS-PROT (Bairoch, et al., 2000, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Res.*, **28**, 45-48) database. The distributions for the five vector components were not uniform, as illustrated for E^1 to E^4 in Figures 10-E1 to 10-E2. For example, the distribution for the component E^1, which correlates well with most hydrophobicity scales (Venkatarajan, et al., 2001, New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties, *J Mol. Model.*, **7**, 445-453), has the most populated bins at the extreme positive and negative values. If the E^1 values in a given column of a multiple alignment are concentrated in a narrow range, especially towards the middle range, the distribution of E^1 values would differ from the 'a priori' distribution and a high relative entropy value is calculated. A physical interpretation is that the residue hydrophobicity is constrained at that position of the protein family during evolution.

In contrast, the component values for E^2, which correlates best with the size/molecular weight of the residues side chains, has a different distribution, with most residues concentrated in the third bin ($5.2 \geq E^2 > 0.2$), and for the fourth vector, which correlates with the natural frequency of occurrence of amino acids (codon degeneracy), bin occupancy decreases as the value of the component E^4 increases. A more detailed discussion on the physical interpretation and importance of the vector components E^1-E^5 is available elsewhere (Venkatarajan, et al., 2001,).

Motifs in the APE family

The APE1 protein family consists of apurinic/apyrimidinic endonucleases and exonucleases. The MOTIFMAKER subroutine of the "PCPMer" program according to the present invention as described with reference to Figure 5 identified 12 motifs of various lengths with the cutoff values for entropy, $\Re$=1.25, for insignificant positions within a motif, G=2, and minimum length, L=4 (See Figure 9A). Most of the motifs are located in the β-strands in the core of the protein. All residues known to be involved in metal ion binding of APE1, including 68N, 96E, 210D, 212N, 308D, and 309H are part of the twelve motifs. The three PROSITE motifs (Falquet, et al., 2002, The PROSITE database, its status in 2002, *Nucleic Acids Res.*, **30**, 235-238) defined for APE correspond to the motifs 2, 9, 10 and 11 defined by the present invention.

Figure 11 shows a qualitative description of a motif, where + or – indicates significant components at the given position with the average values and * means an insignificant component.

With respect to Figure 9A, the 12 motifs of the APE family identified according to the present invention, were used as query sequences to locate the best scoring windows in the sequences of human APE (1hd7.pdb; S1), *E. coli* exonuclease (1ako.pdb; S2), bovine Dnase I (2dnj.pdb, S3) and the IPPP domain of synaptojanin from yeast (1i9y.pdb, S4). These S values are compared to the distribution of values for the highest scoring window over all sequences in the APE family belonging to APE and in the ASTRAL40 database. The * indicates structurally equivalent motifs, based on FSSP/DALI alignments for the four PDB files.

The 12 motifs are differentially conserved in the APE family

Scanning the individual APE sequences with the motif profiles identifies most of the motifs in an equivalent position as in the multiple sequence alignment. Motifs 1, 2, 7, 11 and 12 are particularly well conserved and have consistently high score (S) values. However, motifs 4, 6 and 10 score lower in enteric pathogenic bacteria such as *H. influenza, E. coli, S. enterica, Y. pestis* and *V. cholerae*, which might be related to the observed differences in the activity compared to eukaryotes (Schein, et al., 2002) (see dendrogram of the APE family in Figure 14). Figure 14 shows a dendrogram of 42 different apuinic/apyrimidinic endonucleases and exonucleases used for the multiple alignment of the APE family.

Number(s) indicate motif(s) that failed to score the highest among competing scores from the rest of the scanning windows.

Distribution of scores in ASTRAL40

Figures 12A-12L compares distribution of the highest scores for all motifs in each of the 3635 sequences in the ASTRAL40 database (solid line) to the target scores in the APE family (dashed line). Both distributions can be approximated as a Gaussian function. The scores for motifs 1, 2, 3, 5, 7, 11 and 12, most specific to the APE family, are clearly distinguished.

Ranking all proteins in ASTRAL40 according to the overall score for APE similarity, it was found that all the known members of the DNase-like superfamily at the top of the ranking (See Figure 9B). The Figure indicates that close homologues should have an overall score greater than about 0.5. All sequences with scores between 0.35 and 0.5 were either phosphatases and/or contained a metal ion binding site. Two of the proteins (1MDA and 1EKM) have catalytic centers containing Cu(II), one has 2 Zn^{2+} ions (1QQ9), and another contains Fe(II) (1MPY).

Figure 9B shows a table of the APE related sequences in the ASTRAL40 database. Motifs that scored higher than their average scores in the database were considered as hits and the sequences were ranked according to the bit score obtained for all motif hits.

Identification of APE protein family and superfamily using Molegos

If a 3D structure is known for one or more proteins in the aligned sequences or protein family, the sequence and 3D structural motifs can be combined to find related proteins (Schein, et al., 2002). Structurally related motifs, or molegos, were defined as those segments with a fractional window score greater than 0.6 and a RMSD value less than 2.5 Å for C atoms (See Figure 9C). As with the scoring method based only on sequence, the higher ranking sequences with scores >0.5 were known members of the DNase-I/APE/IPP superfamily. Six proteins with no overall structural or sequence similarity to this superfamily had scores between 0.3 and 0.5. Of these, three contained one (1D09) or two Zn^{2+} (1QQ9) and 1ATL) ions in their active site, one (1D2N) contained Mg^{2+}, and one Ca^{2+}.

Figure 9C shows a table of the APE related sequences in the ASTRAL40 database. Motifs scoring with a fractional window score greater than 0.6 and RMSD less than or equal to 2.5Å were considered to be molegos. Sequences were ranked according to the bit score

obtained for all molego hits. Metal ions present in the protein structures are indicated in brackets.

PSI-BLAST and BLOCKS search for APE family and DNase-I superfamily members

Identification of members of a superfamily using the currently available sequence profile methods is difficult. For example, PSI-BLAST searching, using a local program or NCBI web-based version with default parameters (E-value 0.005), detected members of the APE family in the non-redundant sequence database. However, neither version revealed DNase-I or IPP sequences even after several iterations. When the E-value was increased to 0.1, synaptojanin was revealed within the first iterations, but bovine DNase-I was only detected after 4 iterations, along with more than 500 additional entries. PSI-BLAST also failed to recognize DNase-I or synaptojanin in the ASTRAL40 database, even when we added APE sequences to allow it to form a profile. The BLOCKS search engine did not recognize homology to DNase-I even when the E-value cutoff was extended to 100. In contrast, our method identified these proteins clearly in the structural database (see Figures 9B and 9C).

Identifying and using PCP motifs and molegos

A search method has been developed to locate elements with similar physical-chemical motifs in distantly related proteins. It is demonstrated that the PCP motifs generated by the present invention from a multiple alignment of APE sequences correlated well with motifs identified by other methods, and our database search method efficiently located known homologues. Our Bayesian scoring method discriminated members of the DNase-I superfamily from the bulk of sequences in the representative ASTRAL40 database (see Figure 9B). Further, it is shown that combining sequence and structural data effectively discriminates proteins, with no overall similarity that shares partial function (metal binding) with the APE family.

E^1-E^5 vectors provide an alternative scoring method for evaluating homology

All the commonly used methods for genome sequence searching rely on similar, statistically derived scoring matrices (Altschul, et al., 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389-3402; Kostich, et al., 2002, Human members of the eukaryotic protein kinase family, *Genome Biology*, **3**, 43). Frequently, the same scoring matrix is used to search for related sequences

(for example, with BLAST), prepare a multiple alignment to analyze sequence conservation and to locate distant relatives of the family according to motif conservation. According to, Venkatarajan, et al., 2001, New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties, *J. Mol. Model.*, **7**, 445-453, the five property vectors represent all known physical-chemical properties, and provide an alternative to using the amino acid alphabet (Rigoutsos, et al., 2002, Dictionary-driven protein annotation, *Nucleic Acids Res.*, **30**, 3901-3916) or selected physical-chemical properties (Dubchak, et al., 1999, Recognition of a protein fold in the context of the SCOP classification, *Proteins*, **35**, 401-407) to identify homology. Our PCP motifs complement existing methods for functional cross networking of protein families (Marcotte, et al., 1999, A combined algorithm for genome-wide prediction of protein function, *Nature*, **402**, 83-86; Marcotte, E.M., 2000, Computational genetics: finding protein function by nonhomology methods, *Curr. Opin. Struct. Biol.*, **10**, 359-365; Overbeek, et al., 1999, The use of gene clusters to infer functional coupling, *Proc. Natl Acad. Sci. USA*, **96**, 2896-2901).

Shared PCP motifs and molegos identify DNase-I superfamily members

Despite their low overall sequence identity, APE, DNase-I and IPP families share a similar 3D structure and are members of a superfamily with a common SCOP designation (Lo Conte, et al., 2002, SCOP database in 2002: refinements accommodate structural genomics, *Nucleic Acids Res.*, **30**, 264-267). The present invention rapidly identified members of this superfamily based on their shared PCP motifs. The most conserved motifs, 1, 2, 7 and 12, were found in structurally equivalent regions as defined by FSSP/DALI (Holm, et al., 1996, Mapping the protein universe, *Science*, **273**, 595-602) in an alignment for the DNase-I superfamily members. The structurally equivalent motifs, or molegos, identified by our program are boxed in the FSSP alignment of DNase-I superfamily in Figure 13. These motifs dictate the formation of a β-strand core that serves as the supporting architecture for metal ion binding and phosphorolysis by members of the APE/DNase-I/IPP superfamily (Schein, et al., 2002). No non-APE protein sequence in the ASTRAL40 database scored higher than the average total score (S_x) of 1772 bits calculated for the 42 APE sequences. Thus the present invention is highly sensitive and specific for detecting APE related sequences.

Proteins with highest scores relative to APE bind metal ions

A high proportion of metal ion binding proteins were found as the highest scoring proteins by searching for proteins that have similar sequences and 3D structure motifs (See Figure 9C). Figure 9C indicates that all of the highest scoring proteins identified in the ASTRAL40 structural database, that is those that contained molegos most similar to those of the APE superfamily, use metal ion based catalysis.

This result demonstrates that a new automated method for identifying protein motifs that are conserved according to physical-chemical properties in aligned sequences is provided. PCP profiles of these motifs can be used to locate distantly related proteins in sequence database. The 12 motifs identified according to the methods of the present invention in the APE family include the signatures in the PROSITE database and all amino acids shown previously to be essential for function. The motif profiles successfully identified likely homologues of APE in a database, including several with no overall sequence or structural similarity. Further, combining sequence and structural data to locate proteins that share functional similarities has also been shown.

All references cited herein are incorporated in their entirety as if each were incorporated separately. This invention has been described with reference to illustrative embodiments and is not meant to be construed in a limiting sense. For example, when the term column is used herein, a column could be row if a translation of other terms in accordance therewith is performed. This is the same for other terms, such as, vertical, horizontal or any other directional element. These terms are used to simplify the understanding of the present invention and not to be unduly limiting on the scope of the present invention. Various modifications of the illustrative embodiments, as well as additional embodiments of the invention, will be apparent to persons skilled in the art upon reference to this description.